

# **STAT 419 Course Project: NHANES**

## **Group 5**

Alex Ralston      Delaney Siegmund      Kaviya Veerasingam  
Libby Brill      Tara Rajagopalan

## Section A: Introduction/Background

The National Health and Nutritional Examination Survey provided information for this data set. The information was collected on adults who are at an age of 20 years or older between the years of 1988 and 1994.

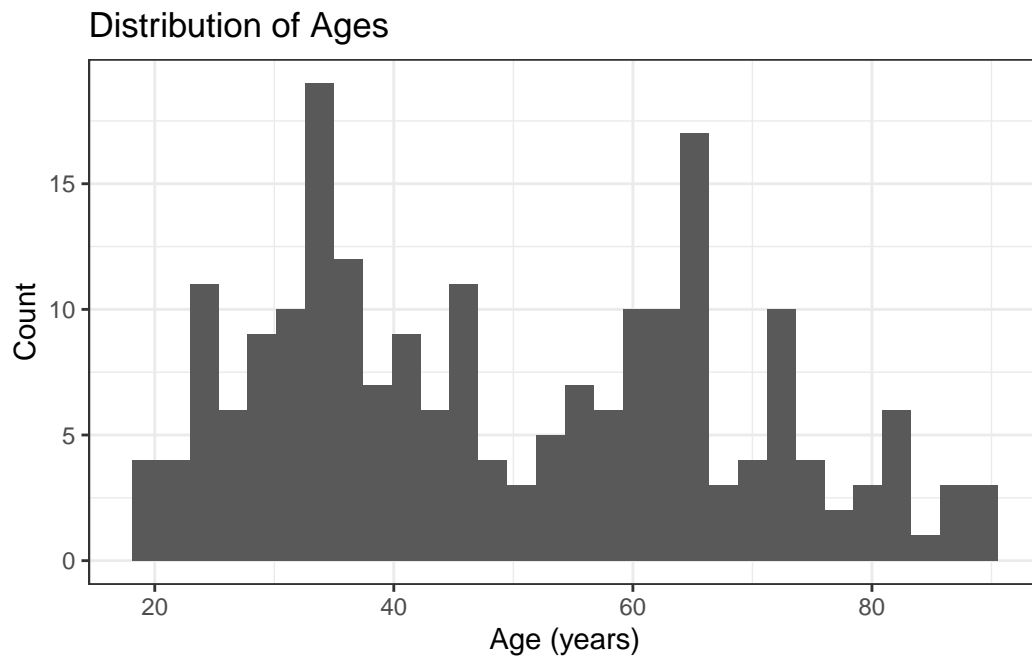
The goal of this data set is to provide the Blood Pressure Rank, measured by variable SBPRANK, for each subject in the Data Set. SBPRank is the grouping variable that is determined by average systolic blood pressure (AVGSBP). The values for SBPRank include 1,2, and 3. 1 is the lowest blood pressure rank indicated by low values of AVGSBP and 3 is the highest blood pressure rank indicated by high values of AVGSBP. 2 is the blood pressure ranking for average systolic blood pressure in the middle ranges.

The other variables in the data set includes:

- HSAGEIR: Subject age recorded in years.
- BMPWTLBS: Subject body weight in pounds.
- BMPHTIN: Subject standing height in inches.
- PEPMNK5R: Subject average diastolic blood pressure
- TCP: This is a measure of the Subject Serum Cholesterol.

## Section B: Graphs and Summary Statistics

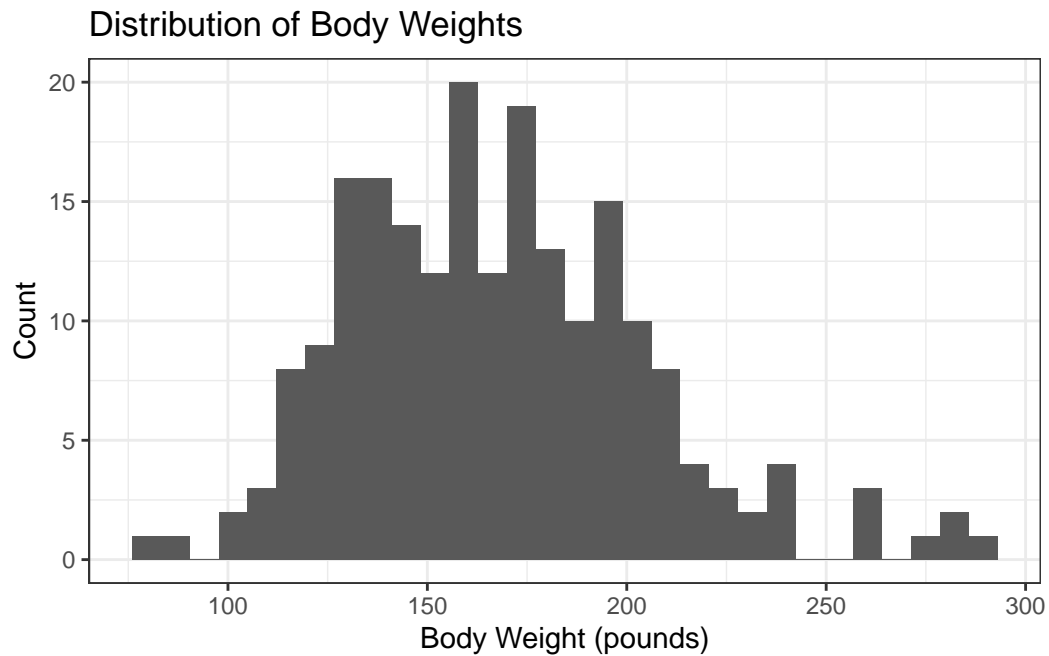
HSAGEIR



	Mean	Median	Standard.Deviation
1	49.44019	46	18.51993

With a mean of 49.44 years, a median of 46 years, and a standard deviation of 18.52 years, the ages in the `nhanes` dataset range from 20 to 90 years old. The distribution of ages appears to be slightly right-skewed and bimodal, with peaks around 35 and 65 years.

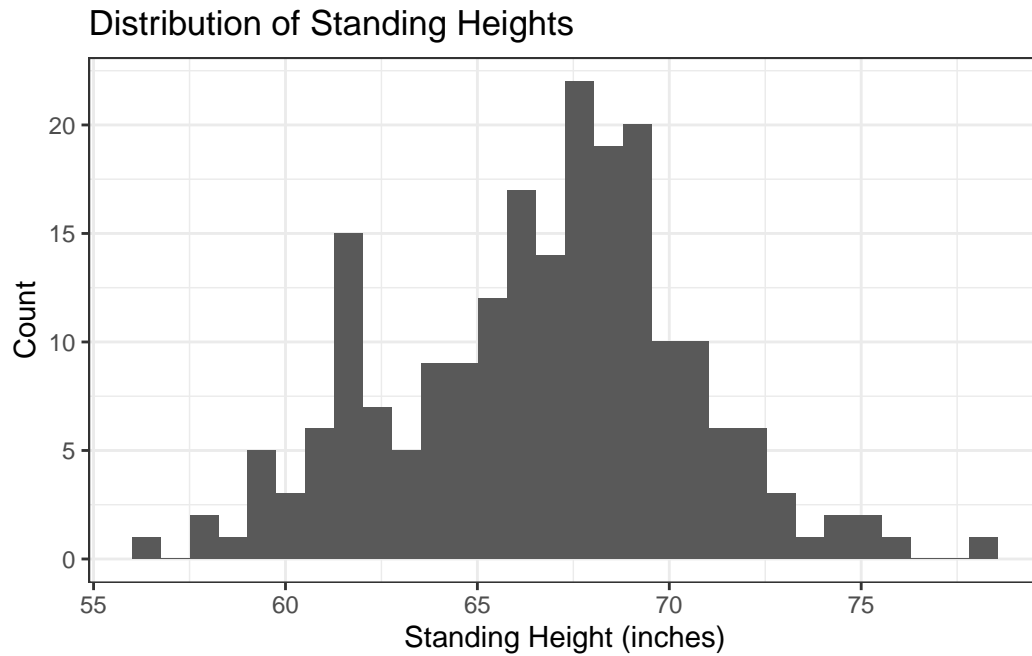
## BMPWTLBS



	Mean	Median	Standard.Deviation
1	167.5651	163	37.55311

With a mean of 167.57 pounds, a median of 163 pounds, and a standard deviation of 37.55 pounds, the body weights in the `nhanes` dataset range from roughly 80 to 300 pounds. The distribution of body weights appears to be unimodal and slightly right-skewed, with some potential outliers in both directions.

## BMPHTIN

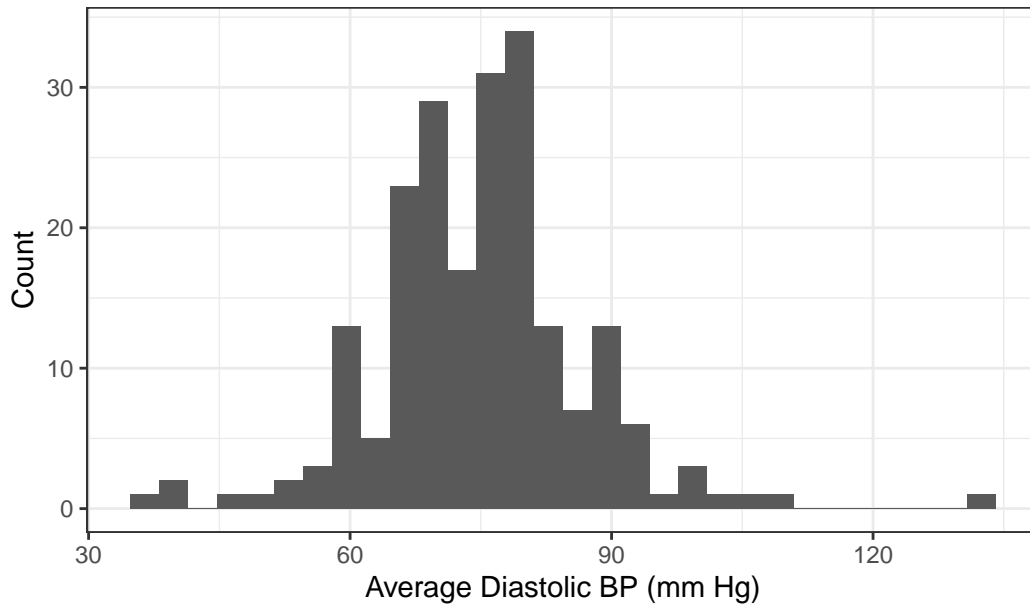


	Mean	Median	Standard.Deviation
1	66.6933	67.2	3.816759

With a mean of 66.69 inches, a median of 67.2 inches, and a standard deviation of 3.82 inches, the standing heights in the **nhanes** dataset range from roughly 55 to 80 inches. The distribution of standing heights appears to be unimodal and fairly symmetric (perhaps slightly left-skewed), with some potential higher outliers.

## PEPMNK5R

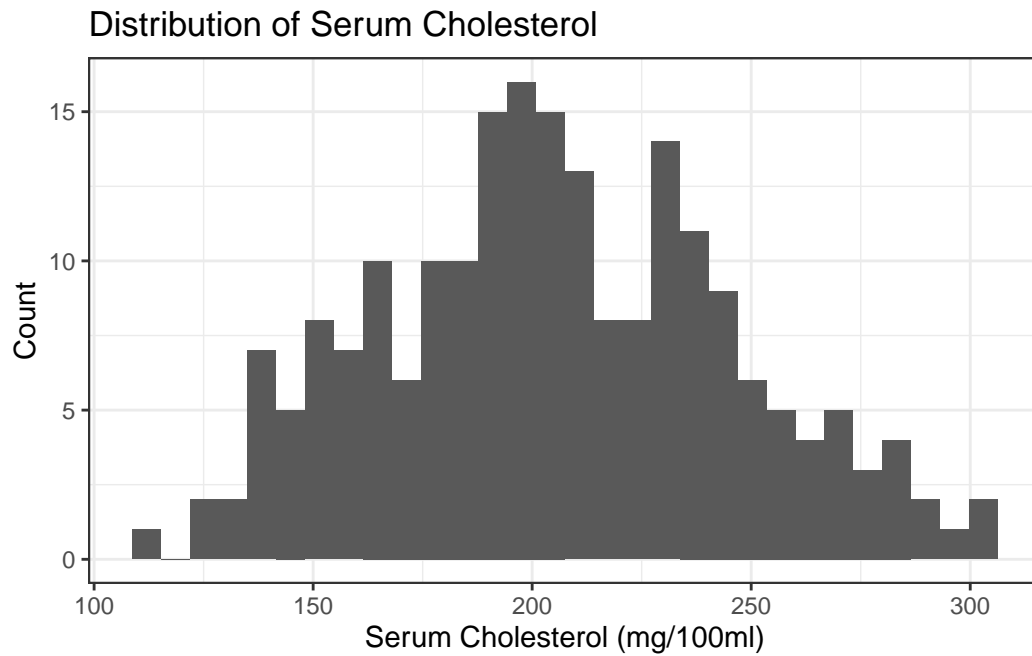
Distribution of Average Diastolic BP's



	Mean	Median	Standard.Deviation
1	74.90909	75	11.79498

With a mean of 74.91 mm Hg, a median of 75 mg Hg, and a standard deviation of 11.79, the average diastolic bp in the `nhanes` dataset range from roughly 35 to 135 mm Hg. The distribution of average diastolic bps appears to be unimodal and fairly symmetric, with some clear outliers in the 130's.

## TCP



	Mean	Median	Standard.Deviation
1	205.6268	203	40.30144

With a mean of 205.63 mg/100ml, a median of 203 mg/100ml, and a standard deviation of 40.30, the serum cholesterols in the `nhanes` dataset range from roughly 110 to 300 mg/100ml. The distribution of serum cholesterols appears to be roughly unimodal (with a slight drop in the 200's) and slightly right-skewed, with no clear outliers.

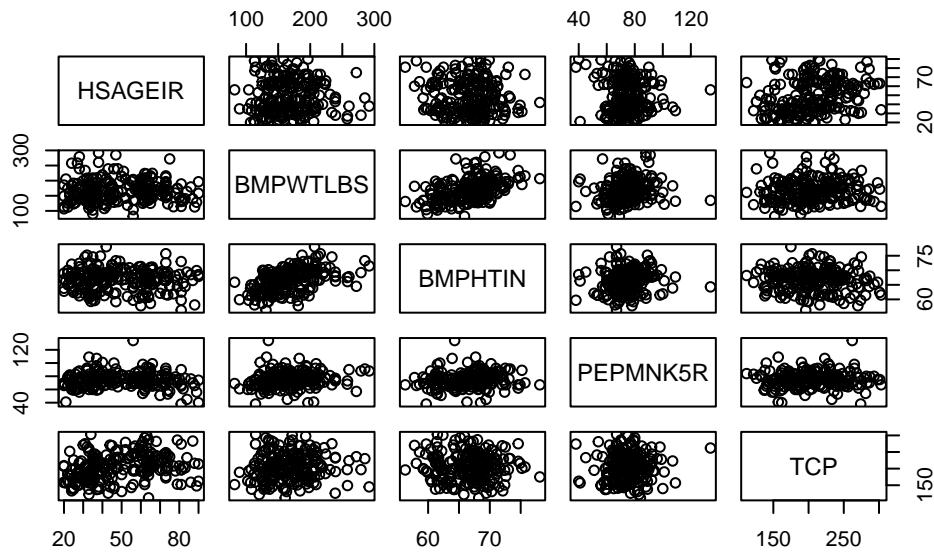
## Section C: Discriminant and Classification Analysis

### C.1: Correlated Quantitative Variables (Multicollinearity)

The Correlation Coefficient Tables for between quantitative variables in the NHANES data set are below:

	HSAGEIR	BMPWTLBS	BMPHTIN	PEPMNK5R	TCP
HSAGEIR	1.000000000	-0.008383008	-0.06742182	-0.0729738	0.28193987
BMPWTLBS	-0.008383008	1.000000000	0.48165357	0.1720942	0.08077624
BMPHTIN	-0.067421821	0.481653572	1.00000000	0.1387325	-0.06180148
PEPMNK5R	-0.072973801	0.172094241	0.13873253	1.0000000	0.04779740
TCP	0.281939870	0.080776245	-0.06180148	0.0477974	1.00000000

The Scatter Plots representing correlations between the quantitative variables in the NHANES data set are below:





## C.2: Discriminant Analysis

	[,1]	[,2]
[1,]	-10.784324	8.677779
[2,]	-3.772346	-7.113163
[3,]	-7.754878	-8.769770
[4,]	2.881487	-2.341885

Let:

- $y_1$  = HSAGEIR (Age) (variable 2)
- $y_2$  = BMPWTLBS (Body Weight) (variable 3)
- $y_3$  = PEPMNK5R (Average Diastolic BP) (variable 5)
- $y_4$  = TCP (Serum Cholesterol) (variable 6)

Then the standardized discriminant functions are:

$$LD_1(y) = -10.784y_1 - 3.772y_2 - 7.755y_3 + 2.881y_4$$

$$LD_2(y) = -8.678y_1 + 7.113y_2 + 8.770y_3 + 2.342y_4$$

We now rank the standardized coefficients of each discriminant function,  $LD_1$  and  $LD_2$ , by observing their absolute values.

For the first discriminant function ( $LD_1$ ),  $y_1$  (age) and  $y_3$  (average diastolic BP) are the most important for separating the groups, followed by  $y_2$  (body weight) and  $y_4$  (cholesterol):  $y_1 \rightarrow y_3 \rightarrow y_2 \rightarrow y_4$ .

For the second discriminant function ( $LD_2$ ),  $y_3$ ,  $y_1$ , and  $y_2$  are the most important, followed by  $y_4$ :  $y_3 \rightarrow y_1 \rightarrow y_2 \rightarrow y_4$ .

### Significance Tests for Discriminant Functions

	Lambda	V	p.values
LD1	0.4856068	147.721810	0.00000000
LD2	0.9596942	8.413261	0.03820007

### Hypotheses:

- 1st Test (for  $LD_1$ ):
  - $H_0$ :  $\alpha_1 = \alpha_2 = 0$
  - $H_a$ : At least one  $\alpha_i \neq 0$

- 2nd Test (for  $LD_2$ ):
  - $H_0: \alpha_2 = 0$
  - $H_a: \alpha_2 \neq 0$

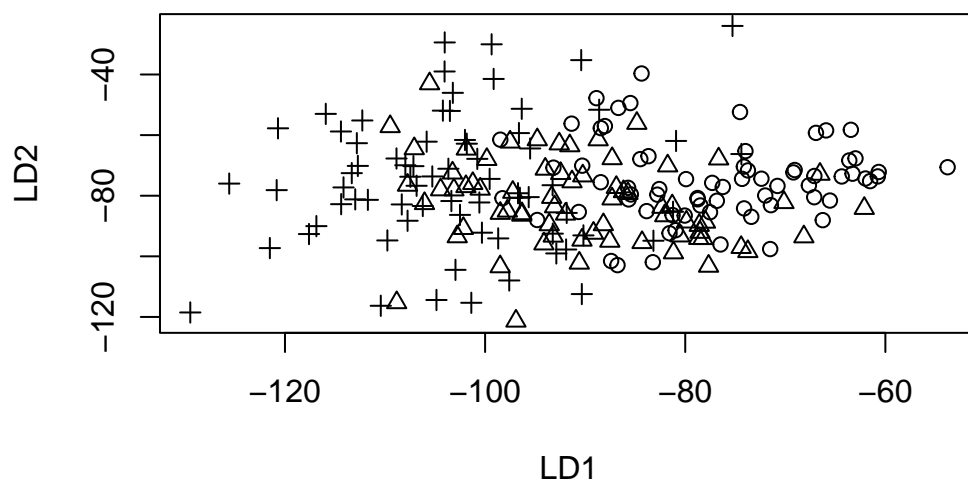
### Conclusions:

- blabla..

### Significance Tests for Additional Variables

	Lambda	F.stat	p.value
HSAGEIR	0.6046790	66.357652	0.000000e+00
PEPMNK5R	0.7659335	31.018029	1.759481e-12
BMPWTLBS	0.9344024	7.125580	1.021397e-03
TCP	0.9679940	3.356017	3.681948e-02

### Visualizing Discriminant Functions



### C.3: Classification Analysis

#### Four most significant variables

#### Classification Analysis

```
$coefs
      [,1]      [,2]      [,3]      [,4]
[1,] 0.2916651 0.1120130 0.6792672 0.07939680
[2,] 0.3555073 0.1287796 0.7809797 0.07280753
[3,] 0.4457149 0.1350911 0.8384779 0.06346110
```

```
$c.0
[1] -45.52791 -57.05103 -65.63240
```

Given:

- $y_1$  = Age (Variable 2)
- $y_2$  = Body Weight (Variable 3)
- $y_3$  = Average Diastolic BP (Variable 5)
- $y_4$  = Serum Cholesterol (Variable 6)

Then:

$$L_1(y) = 0.292y_1 + 0.112y_2 + 0.679y_3 + 0.079y_4 - 45.528$$

$$L_2(y) = 0.356y_1 + 0.129y_2 + 0.781y_3 + 0.073y_4 - 57.051$$

$$L_3(y) = 0.446y_1 + 0.135y_2 + 0.838y_3 + 0.063y_4 - 65.632$$

Assign to  $G_1$ ,  $G_2$ , or  $G_3$  depending on if  $L_1(y)$ ,  $L_2(y)$ , or  $L_3(y)$  respectively yields the greatest value for a given observation.

#### Apply Classification to Observation 1

Observation 1, eliminating height:

```
nhanes(1)' = (63 141.4 64 20)
```

```
      [,1]
[1,] 49.62603
```

```
      [,1]
[1,] 49.55572
```

```

      [,1]
[1,] 49.17355

```

For a single observation, the linear classification function with the greatest returned value is considered to be assigned to that group. For the first observation, the first linear classification function returned the greatest value of 49.626, indicating that the observation be correctly classified as part of Group 1.

This is very similar to the values of other functions  $L_2(y)$  and  $L_3(y)$  (49.556 and 49.174 respectively), which could indicate a large amount of noise in the data. Non-linear classification functions might explain separation of groups better if further analysis were to be conducted.

### Confusion Matrix

```

$`Correct Class Rate`
[1] 0.6363636

```

```

$`Error Rate`
[1] 0.3636364

```

```

$Method
[1] "LDA"

```

```

$`Confusion Matrix`
      predicted
original 1  2  3
      1 51 15  3
      2 18 32 20
      3  5 15 50

```

ACCR = 63.64%

AER = 36.36%

The linear classification functions are classifying the majority of observations correctly (64%), but there is some classification error (36%). This supports earlier conclusions of high noise within the data leading to some incorrectly classified observations.

## Section D: Summary

## Section E: R Code

### Setup

```
library(tidyverse)
source("all_customized_functions.R")
nhanes <- read.csv(here::here("NHANES3_419.csv"), header = TRUE)
```

### Section B

```
# HSAGEIR
nhanes |>
  ggplot(aes(x = HSAGEIR)) +
  geom_histogram() +
  labs(x = "Age (years)", y = "Count", title = "Distribution of Ages") +
  theme_bw()

data.frame(Mean = mean(nhanes$HSAGEIR), Median = median(nhanes$HSAGEIR),
            `Standard Deviation` = sd(nhanes$HSAGEIR))

# BMPWTLBS
nhanes |>
  ggplot(aes(x = BMPWTLBS)) +
  geom_histogram() +
  labs(x = "Body Weight (pounds)", y = "Count",
       title = "Distribution of Body Weights") +
  theme_bw()

data.frame(Mean = mean(nhanes$BMPWTLBS), Median = median(nhanes$BMPWTLBS),
            `Standard Deviation` = sd(nhanes$BMPWTLBS))

# BMPHTIN
nhanes |>
  ggplot(aes(x = BMPHTIN)) +
  geom_histogram() +
  labs(x = "Standing Height (inches)", y = "Count",
       title = "Distribution of Standing Heights") +
  theme_bw()
```

```

data.frame(Mean = mean(nhanes$BMPHTIN), Median = median(nhanes$BMPHTIN),
            `Standard Deviation` = sd(nhanes$BMPHTIN))

# PEPMNK5R
nhanes |>
  ggplot(aes(x = PEPMNK5R)) +
  geom_histogram() +
  labs(x = "Average Diastolic BP (mm Hg)", y = "Count",
       title = "Distribution of Average Diastolic BP's") +
  theme_bw()

data.frame(Mean = mean(nhanes$PEPMNK5R), Median = median(nhanes$PEPMNK5R),
            `Standard Deviation` = sd(nhanes$PEPMNK5R))

# TCP
nhanes |>
  ggplot(aes(x = TCP)) +
  geom_histogram() +
  labs(x = "Serum Cholesterol (mg/100ml)", y = "Count",
       title = "Distribution of Serum Cholesterol") +
  theme_bw()

data.frame(Mean = mean(nhanes$TCP), Median = median(nhanes$TCP),
            `Standard Deviation` = sd(nhanes$TCP))

```

## Section C1

```

NHANESData <- nhanes |>
  select(-1)

# correlation table
cor(NHANESData)

# scatter plot
plot(NHANESData)

```

## Section C2

```
nhanes_c2 <- nhanes
nhanes_c2$SBPRANK <- as.factor(nhanes$SBPRANK)
# Omit height variable
nhanes_c2 <- nhanes_c2[, c("SBPRANK", "HSAGEIR", "BMPWTLBS",
                           "PEPMNK5R", "TCP")]

X <- nhanes_c2[, -1]
y <- nhanes_c2[, 1]

# discriminant analysis
discrim(X, y)$a.stand

# significance tests for discriminant functions
discr.sig(X, y)

# significance tests for additional variables
partial.F(X, y)

# visualizing discriminant functions
discr.plot(X, y)
```

## Section C3

```
nhanes_c3 <- nhanes
# Omit height
data <- nhanes_c3[, -4]

# classification analysis
lin.class(data[, -1], data[, 1])

# apply classification to observation 1
L1 <- 0.2916651%*(63) + 0.1120130%*(141.4) + 0.6792672%*(64)
+ 0.07939680%*(220) - 45.52791
L2 <- 0.3555073%*(63) + 0.1287796%*(141.4) + 0.7809797%*(64)
+ 0.07280753%*(220) - 57.05103
L3 <- 0.4457149%*(63) + 0.1350911%*(141.4) + 0.8384779%*(64)
+ 0.0634611%*(220) - 65.63240
```



```
L1  
L2  
L3  
  
# confusion matrix  
rates(data[,-1], data[,1])
```